Artificial intelligence method to design and fold alpha-helix structural proteins from the primary amino acid sequence

Zhao Qin[1‡], Lingfei Wu[2,3‡], Hui Sun[1‡], Siyu Huo[3], Tengfei Ma[3], Eugene Lim[1], Pin-Yu Chen[2,3], Benedetto Marelli[1,2*], and Markus J. Buehler[1,2*]

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave. 1-290, Cambridge, MA 02139, United States of America

[2] MIT-IBM Watson AI Lab, 75 Binney St, Cambridge, MA 02142, United States of America

[3] IBM Research AI, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, United States of America

[‡] These authors contributed equally to this work

[*] Corresponding authors: Benedetto Marelli, bmarelli@mit.edu, Markus J. Buehler, mbuehler@mit.edu

Abstract: We report an artificial intelligence (AI) based method to predict the molecular structure of proteins, focused here on an important subclass of proteins dominated by alpha-helix secondary structure, as found in many structural biomaterials such as keratin and membrane proteins. Fast yet accurate predictions of an unknown protein's 3D all-atom structure can yield a pre-screened set of candidate proteins to be investigated further via large-scale protein expression in bacteria or yeast. However, classical molecular simulations are greatly limited by the time scale and significant computational cost needed for the complete folding of a long peptide into a complex structure from scratch, which can easily exceed the capability of a supercomputer. To accelerate simulations at low computational cost here we report an innovative machine learning method to offer a high-throughput prediction of the protein structure, as well as the material and biological functions from purely the protein sequences. To achieve this, we designed a novel Multi-scale Neighborhood-based Neural Network (MNNN) model that is capable of learning the neighborhood structured information in the raw protein sequence trained on the database of over 120,000 protein structures. The method directly predicts the phi-psi dihedral angles of the backbone of each constituting amino acid, which is then used to construct the full all-atom 3D structure of the corresponding protein without any template or co-evolutional information. We find that our machine learning model can accurately predict all dihedral angles of any target sequence. The prediction yields a maximum average error of 2.1 Å of the predicted 3D structure compared with experimental measurement. We find that the predicted folded structure from MNNN consumes less than six orders of magnitude time than classical molecular dynamics simulations, offering extremely fast folding predictions. Our results suggest that the MNNN model can be used to greatly accelerate the prediction of protein structures.

Keywords: Protein; artificial intelligence; machine learning; deep neural networks; folding; structure prediction; computation

Introduction

The development of rational techniques to discover new proteins for use in variety of applications ranging from agriculture to biotechnology remains an outstanding materials design problem[1,2]. In fact, proteins represent the key construction materials of the living world, and offer enormous diversity in function, and hence a powerful platform for potential for use in bioengineering, medicine and materials science. Among several universal secondary structures, alpha-helices (AHs) are a universal motif of many biological protein materials. These protein domains play a crucial role in the signaling and deformation behavior of cytoskeletal protein networks in cells (e.g. intermediate filaments, as well as actin) [3,4], and in determining the mechanical properties of hair, hoof, feather and many other important structural protein materials (e.g., keratin) [5]. Several silk proteins also possess helical structures, which impart toughness and antimicrobial properties to the final material [6–8]. AHs are also the most common structural motif in cellular membrane proteins, which are responsible for transport of matter, cell recognition, docking and signal transduction. In these materials, nanostructured AH based protein domains universally define their nanoscale architecture. Although the Protein Data Bank (PDB) [9] provides a rich resource of folded protein structures and their all-atom 3D geometries (~120,000 protein structures to date), this database only includes a tiny portion of all proteins known to exist. Most proteins, however, are only known by their sequence and a limited set of associated functions (such as the 147,413,762 protein sequences given in uniprot.org [10]), and their high-resolution protein structure remains unknown. Indeed, it is difficult to identify the complex structure of a protein from a pure experiment, which requires advanced tools including Nuclear Magnetic Resonance, X-Ray Diffraction or Cryo-electron microscopy, as well as protein crystal samples. Many proteins cannot be investigated in that way and hence, their full 3D structure, full set of functions (including roles in disease etiology or as platform for biomaterials) remains elusive.

The protein dynamics revealed by atomistic simulation in an accurate solvent model condition can provide an accurate description of how a protein changes its conformation toward the state with lower free energy. It has been demonstrated that the 3D folded structure of a protein can be obtained from a computational simulation of protein folding directly from the sequence information[11,12]. However, the final equilibrated structure greatly depends on the initial conformation, as the structure can easily be trapped at a local energy minimum, while the global energy minimum can only be reached by crossing energy barriers, which are very rare transition events that must happen during a classical molecular dynamics (MD) simulations. A typical protein of ~100 amino acids could require few seconds to fold. As classical MD computes the interactions and motions of a large number of atoms stepwise and as each time increment must be on the order of 1~2 fs [13], it would require 10^{15} computational integration operations for the full simulation of the folding trajectory, which goes beyond the capability of most supercomputers. Other methods such as the Replica Exchange Method [14] effectively combine different simulation algorithms to greatly accelerate the folding calculation compared with classical MD, but are still not fast enough to provide rapid results.

Artificial intelligence (AI), enabled by deep learning (DL) techniques, has demonstrated its advantage in solving sophisticated scientific problems that involve multiple physics interactions that are challenging to directly model or non-polynomial problems that require extremely large computational power that cannot be solved by brute force [15]. Recent work has suggested that it may provide a feasible way to achieve fast prediction of protein structures by utilizing efficient algorithms of searching a high-dimensional parameter space for the most accurate prediction. In several materials-focused studies, such a data-driven material modeling for optimized mechanical properties of materials, it has shown its great potential in advancing conventional multiscale models in terms of efficiency and speed of predictions [16–19]. However, it is believed that the capability to optimize the multiscale and multiparadigm architecture of materials features a high sensitivity to environmental factors, as needed in sensors, electronics and for multi-purpose material applications [20].

Computational approaches for dihedral angle predictions or direct protein folding can be categorized into two general categories. The first category leverages existing identified protein structure templates [23] or co-evolving residues information within protein families [24] to derive a sequence-to-structure contact map. However, these conventional methods are often limited to analyzing small proteins due to the high computational cost. More importantly, they are unable to predict structures for which no existing protein templates or co-evolution information can be used. The second category attempts to explore more advanced machine learning methods, especially recent deep learning techniques, to build an end-to-end architecture for directly predicting dihedral angles [25] or 3D structures of proteins [22]. However, all deep learning methods reported thus far still heavily rely on domain specific input features beside the primary amino-acid sequence, making it hard to generalize new protein sequences with have no known information about these input features.

In this paper, we present a Multi-scale Neighborhood-based Neural Network (MNNN) model for predicting dihedral angles directly from the sequence neighborhood by taking into account both raw sequence neighborhood and secondary sequence neighborhood. It is worth noting that our model takes only primary raw amino-acid sequence as inputs and directly predicts dihedral angles without any template or co-evolutional information. To fully exploit the information of the secondary structure information, we also propose to use a data-driven approach (clustering techniques such as K-means clustering) to compute the possible number of secondary structure classes instead of the conventional eight-class categorization. Our experimental results show that our MNNN model can accurately predict the complete set of dihedral angles of a target sequence with small prediction errors (see Materials and Methods for details).

The outline of the paper is as follows. We describe the design, training and validation of the proposed MNNN model, and then proceed to apply the model to make predictions of existing proteins and *de novo* sequences. We report a series of all-atom explicit solvent molecular simulations to confirm the stability of the predicted proteins. To demonstrate that the MNNN model can make accurate predictions for *de novo* sequences outside of the training set, we experimentally synthesize a short sequence that is not included in the PDB and fully characterize its structure content in experiment, and compare it with both MNNN and MD predictions. This three-way validation confirms the predictive power of our approach.

Results

Integrated Framework for Protein Design

We report an integrated framework that combines MNNN, MD and experimental protein synthesis for protein designs, as shown in **Fig. 1**. Compared to experience-based trial-and-error for protein synthesis, this framework allows high-throughput *in silico* prediction of protein structures and related material functions that provide a rational basis for the design of *de novo* protein materials. We focus on achieving a capability of fast prediction of alpha-helical proteins because as one of the major universal secondary structural motifs, alpha helices provide a platform for materials design with wide ranging implications in a variety of application areas. For example, experimental studies revealed that α-keratin (found in wool, hair and hooves), fish slime threads, desmin and vimentin are all composed of alpha-helices and are thus stretchable and tough protein materials.

Despite the relatively simple geometry and well-known mechanical functions of alpha-helical proteins, fast predictions of their all-atom 3D folded structure is crucial to identify biological functions of a protein (e.g., binding, docking, assembly into higher-order structures, or specific biological properties such as antimicrobial). To achieve this goal, our MNNN model learns how the phi-psi angles are mapped to specific peptide sequences for all proteins with known 3D structures in PDB. It can then make predictions of the phi-psi angles for any given protein sequence. These predicted angles, combined with the raw sequence information, are then used to build an all-atom 3D folded protein structure that can serve as the input geometry for further refinement through conventional MD simulations (see Materials and Methods for details). This multi-stage process can be used, for instance, to test the predicted protein's thermal stability at room temperature. Our learning and simulation procedure hence provides a robust computational basis that helps to select and verify the most promising sequences that can lead to alpha helix structures.

Neural Network Architecture for Dihedral Angle Prediction

Our deep learning regression model for predicting dihedral angles directly from the sequence neighborhood is based on a data-driven partition of design space of a protein structure, as shown in **Fig. 2**. Comparing to existing template or co-evolutional based machine learning models for protein structure prediction, our model has the capability of directly predicting protein structures solely from primary raw protein amino-acid sequences. Specifically, our model is an end-to-end machine learning system that only requires raw protein sequences as data inputs and produces phi-psi angle prediction as outputs. To fully exploit all hidden structured information in the protein sequence, we take both raw and secondary sequence neighborhood information into consideration. By considering the raw sequence neighborhood, our model is able to learn the correlation among the subsequent continuous amino-acids. Similarly, the secondary sequence information provides important structure information about the raw amino-acid sequence and thus serves as additional constraints for a MNNN model to achieve a better angle prediction. However, the existing partition of the design space of protein structures (e.g., established methods such as DSSP [21]) only considers an eight-class human-engineered categorization, which may or may not be sufficient to characterize the diversity of natural structures.

To address this challenge, we develop a data-driven approach to compute the possible number of secondary structure classes. We use advanced clustering techniques such as K-means clustering on all PDB data with different cluster numbers and verify its degree of matching in comparison with the benchmark PDB structure, as shown in **Figs. 2A and B**. We found that when the class number is set to 256, the error between simulated structure with the benchmark result is reduced to a small level. Since the secondary structure of neighboring amino acids will influence the subsequent secondary structure of the next amino acid, it is important to consider the neighboring K number of secondary structure prediction information when predicting dihedral angles of the next amino acid. When training a MNNN model (**Fig. 2C**), both data embeddings representing the raw amino acid sequence and their secondary structure are incorporated to learn a refined embedding. We then use the refined embeddings of the neighboring amino-acids for phi-psi dihedral angle predictions of a given sequence of amino acids.

Benchmark for Prediction of Protein Structures

For structure predictions, we find that the MNNN model exhibits a significant advantage and enlarged potential over conventional methods, such as all-atom MD simulations for protein folding. Six well-characterized coiled-coil peptides [22–25] were chosen as benchmark proteins to test the efficiency and accuracy of our MNNN model. The root-mean-square deviation (RMSD, see Materials and Methods) of the MNNN-predicted structures is compared to the available structures in the PDB for the six peptides were computed. The outcomes, as summarized in **Table 1**, are compared against protein folding with classical MD simulations with implicit and explicit solvent models. It is shown that the largest error given by MNNN is merely 2.11 Å, which is much better than folding the peptide from a fully extended form in MD, either in implicit (12.9±4.2 times the error) or explicit (10.4±4.6 times the error) solvents. Most importantly, the time needed to obtain the predicted folded structure from MNNN is significantly smaller (less than six orders of magnitude) than classical MD simulations (with the least requirement of computing hardware). This outcome suggests that the MNNN model can offer a powerful way to predict protein structures.

Moreover, it is also important to notice that the structures given by the MNNN model not only agree well with available structures in the PDB, but also yields predictions with good overall thermal stability. As summarized in Table 1 and shown in Fig. 3, the structure predicted by the MNNN model only appear to have a very small thermal fluctuations in all-atom MD simulation (with RMSD<1.8 Å). It is also noted that the implicit solvent model (much less computationally expensive than explicit solvent models), no matter whether they start from the fully extended form or from the structure given by the MNNN model, does not yield good predictions for either folding or structure refinement purposes.

De novo Protein Design and Structural Validation

Besides the six small proteins whose 3D structures are already resolved, we test the accuracy and efficiency of our MNNN algorithm in protein folding prediction on *de novo* proteins whose molecular structure is unknown. To this end, a peptide of 28 amino acids (named AmelF3_+1) extracted from the coiled-coil domain of the Apis mellifera silk protein (AmelF3)[7] was chosen (**Fig. 5A**). The results of the MNNN model developed in this study to predict the structure of AmelF3_+1 are compared with structure homology prediction tools including Optimized Protein Fold Recognition (ORION) and Iterative Threading Assembly Refinement (I-TASSER), as shown in **Fig. 5B**. ORION is a sensitive method based on a profile-profile approach that relies on a good description of the local protein structure to boost distant protein structure predictions, while I-TASSER features a hierarchical approach for protein structure and function prediction that identifies structural templates from the PDB, with full-length atomic models constructed by iterative template-based fragment assembly simulations. The prediction by our MNNN model is an alpha-helix , which overall agrees with the results of the other two methods, both of which also predict alpha-helical proteins. We use the three predicted structures and compare their dynamic behaviors in 100 ns MD simulation in explicit solvent (**Fig. 5B**). It is found that the structure predicted by the MNNN model is more thermodynamically stable than the other two, particularly significantly better than the I-TASSER model, which almost unfolded during the middle of the simulation run. ORION is limited by template availability, and the prediction of the highest scoring structure may not be of the same length as the targeted sequence and one may need to balance between structure integrity and accuracy.

To validate the computational results, the peptide AmelF3_+1 was synthesized and characterized experimentally. The circular dichorism (CD) spectrum of AmelF3_+1 (**Fig. 5C**) shows a major combination of alpha-helical, beta-turns and random coils conformations [26,27], with the relative contents being 57%, 24% and 14%, respectively, as estimated by the CONTINLL program [26,28]. Moreover, potential peptide assembly into higher-order structures was captured by TEM, as shown in **Fig. 5D**, from which we can see that the peptide assembles into either nanofibers of around 10 nm in width and several microns in length or nanoparticles of diameters ranging from 10-35 nm. The secondary structure analysis was also independently confirmed by ATR-FTIR and Raman spectroscopy, the spectra of which are complementary in the Amide I and III bands, with well-established peak assignments for different secondary structures [29–31]. From the FTIR spectrum of AmelF3_+1 (**Fig. 5E**), a major peak at 1656 cm-1 in the Amide I region along with the two peaks at 1323 and 1304 cm-1 in the Amide III region indicate predominant alpha-helical conformations of the peptide. The Raman spectrum of AmelF3_+1 (**Fig. 5F**) gives similar structural information, with two more hidden alpha-helical peaks (*i.e.*, 1295 cm-1 and 1281 cm-1) clearly seen. It is anticipated that with the correct buffer condition, more stable alpha-helical conformation of the AmelF3_+1 peptide can be achieved.

Discussion and Conclusion

In this paper we reported a new approach to accurately and rapidly predict the structure of *de novo* proteins directly from the primary protein sequence. Proteins are the most abundant building blocks of all living things, and their geometry is linked intimately to functional properties. The AI based approach to design new proteins opens the door to generative methods that can complement conventional protein sequence design methods. Future work could expand the method to include other secondary structures, and achieve a broader, more comprehensive structure prediction capacity.

Materials and Methods

Data Preparation for Training and Translating Predictions to 3D All-Atom Protein Structures

We develop a custom script to obtain the phi-psi angle and sequence information of each of 126,732 natural protein structures composed of only standard amino acids that are currently available in the PDB. We wrote our own bash script that allows using DSSP to compute the phi-psi angles of the backbone by reading the 3D structure files that are automatically downloaded from PDB and using open source Unix software to build a highly structuralized database for training (**Fig. 1**). We develop a post-process script to take the (phi, psi) angle as predicted by MNNN to combine with the rest of the geometric parameters given by the intrinsic coordinates within the CHARMM force field to build the all-atom protein structure, which allow to run energy minimization and molecular dynamics to validate the thermal stability of the protein structure.

Neural Network Training and Validation

The multi-scale neighborhood-based neural network model predicts dihedral angles directly from sequence neighborhood, by taking into account both raw sequence neighborhood and secondary sequence neighborhood. Our model consists of three main steps:

A) We compute character embedding for each amino acid (can be done using a variety of popular techniques) and then refine these embeddings with sequence neighborhood information. For instance, one can use long short-term memory (LSTM) or bidirectional LSTM (Bi-LSTM) to compute character embeddings. One can also use other character embedding techniques such as fasttext, glove, word2vec or transformer to compute character embeddings. In this paper, we choose to first use Bi-LSTM layer to compute character embeddings of primary amino-acid. Then we apply a convolutional neural network (CNN) layer to take into account sequence neighborhood information based on computed character embeddings of each amino acid.

B) Based on refined character embeddings of amino acid sequence, we apply another LSTM layer to perform dihedral angles prediction. For the first amino acid, it uses only its own hidden representation to predict dihedral angles. For the second and subsequent amino acids, we augment the embedding from the previous K number of embeddings of predicted secondary structure character and then predict dihedral angles.

C) Once dihedral angles are predicted, we use these two embeddings to further predict secondary structure characters as additional constraints. In conventional models there are typically eight different secondary structure characters in total. However, this choice is ad-hoc and there are large sets of structural features that are not well covered by this categorization. Therefore, we propose to use a data-driven approach (clustering techniques such as k-means clustering) to compute the possible number of secondary structure classifications.

Our overall loss function for our MNNN model consists of three parts: i) a loss function with RMSE of the predicted Phi angle and the real ones; ii) a loss function with RMSE of the predicted Psi angle and the real ones; iii) a loss function for matching the predicted secondary structure class with real classes (from data-driven approach). Note that the angles are represented by trigonometric functions (*i.e.*, an angle is converted to a pair of its sine and cosine values for numerical stability during model training). We then split the PDB data into train/validation/test datasets and train the model with train/validation data and test our models with test data. We further compute the L1 norm of dihedral angles of any target sequence within average L1-norm errors for the phi-psi angle pair. Throughout our experiment, we set the neighborhood parameter in our MNNN model to be 21. On the test dataset, we obtain average L1 errors of 22 degrees and 37 degrees for the Phi-Psi angles respectively. The training error is similar to the test error.

Molecular modeling

We use two force fields (CHARMM19 with implicit solvent [32], and CHARMM27 with explicit solvent [33]), with each of them starting from two different extreme configurations (full extended chain, and a structure with phi-psi angles predicted from MNNN), to perform individual long-time simulations (100 ns) for each of the protein sequence. During the simulation, the change of the molecular conformation is benchmarked by quantitatively comparing with the corresponding protein structure within the PDB.

$$RMSD(t) = \sqrt{\frac{\sum_{i=0}^{n}\left[\left(x_{i_{MD}}(t)-x_{i_{PDB}}\right)^2+\left(y_{i_{MD}}(t)-y_{i_{PDB}}\right)^2+\left(z_{i_{MD}}(t)-z_{i_{PDB}}\right)^2\right]}{n}} \tag{1}$$

Where *n* is the number of all the backbone atoms of each amino acid of the peptide (N, C, CA, O) and $(x_{i_{MD}}(t), y_{i_{MD}}(t), z_{i_{MD}}(t))$ is the Cartesian coordinates of the backbone atoms given by the MD simulation at time *t*, while $(x_{i_{PDB}}, y_{i_{PDB}}, z_{i_{PDB}})$ is the Cartesian coordinates of the backbone atoms of the protein structure within the PDB. In the CHARMM model, the mathematical formulation for the empirical energy function has the form:

$$E = \sum E_{bond} + \sum E_{angle} + \sum E_{dihedral} + \sum E_{improper} + \sum E_{\mathrm{Urey-Bradley}} + \sum E_{nonbonded} \tag{2}$$

Each energy term is given by $E_{bond} = K_{ij}(r - r_0)^2$ is the bond term that defines how two covalently bonded atoms interact in the stretching direction, $E_{angle} = K_{ijk}(\theta - \theta_0)^2$ is the angle term that defines how the angel among three covalently bonded atoms with one central atom changes under external force, $E_{dihedral} = K_{ijkl}[1 + \cos(n\phi - \delta)]$ is the dihedral term that defines how the dihedral angel among four covalently bonded atoms with one central bond changes under external force, $E_{improper} = K_{ijkl}(\omega - \omega_0)^2$ is the improper angle term that defines how the improper angle among four covalently bonded atoms with one central atom changes under external force, $E_{\mathrm{Urey-Bradley}} = K_u(u - u_0)^2$ is the Urey Bradley term that accounts for angle bending and $E_{nonbonded} = \epsilon_{lj}\left[\left(\frac{R}{r_{ij}}\right)^{12} - \left(\frac{R}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{r_{ij}\epsilon}$ is the nonbonded term that accounts for the van Der Waals (VDW) energy and electrostatic energy. In all-atom force fields, water molecules generally can be treated either explicitly or implicitly for MD simulations.

We use the CHARMM19 all-atom force field to model the atomic interactions for the straight chain and the MNNN predicted model. The solvent effect for this force field is generally considered by using the implicit Gaussian model (EEF1) for the water solvent[32]. The use of the implicit solvent model has advantages to accelerate the sampling speed of molecular configurations. We use the CHARMM c37b1 package to run the simulation for energy minimization and structural equilibration. Because there is no explicit water or pressure control, we do not apply any constraint to ensure the simulation stability. The time step used for implicit solvent simulations is 1 fs.

Starting from the initial geometry built using the backbone dihedrals (with (phi, psi)=(180°, 180°) for each amino acid for a straight chain, and (phi, psi) defined by MNNN for another model), combining with the rest of the geometric parameters given by the intrinsic coordinates within the CHARMM force field, we follow the following protocol to equilibrate the structure: 1) Energy minimization (2,000 Steepest Descent steps followed by 2,000 Adopted Basis Newton-Raphson steps); 2) Equilibration runs for 50 ps (NVT ensemble with Nose-Hoover temperature control), where the temperature rises linearly from 240 K (beginning) to 300 K (end); 3) Equilibration runs for 100 ns (NVT ensemble with Nose-Hoover temperature control), where the temperature stays at 300 K. We record the coordinates for each 10 ps, compare with the PDB structure by RMSD to measure how far is the folded structure away from the PDB structure.

Besides implicit solvent, we use the CHARMM27 force field implemented by the explicit TIP3P water model and run simulation with NAMD package v2.13[34] with the support of Graphical Processing Units (GPUs), which greatly outperforms Central Processing Unit (CPU) performance. Our model starts from the equilibrated structure obtained by using the implicit solvent model described above. All simulations run in a NPT ensemble under a constant temperature (300 K) and constant pressure (1 atmosphere) controlled by Langevin thermostat and barostat. The simulation time step is 2 fs with rigid bonds model for all the covalent bonds between hydrogen atoms and other heavy atoms. We use particle mesh ewald (PME) function with a grid width <1 Å to calculate the electronic interaction because and it is an efficient method to accurately include all the long-distance electrostatics interactions. A cutoff length of 10 Å is applied to the van der Waals interactions and to switch the electrostatic forces from short range to long-range forces.

The initial protein structure for explicit solvent simulation is built the same way as the implicit models. We use Visual Molecular Dynamics (VMD) [35] to add a solvent box around the protein structure with water at a distance of least 10 Å from the protein structure. The net charge of the system is set to zero by adding NaCl of overall concentration of 0.1 Mol/L, and each ion is initially randomly placed in the solvent box with the actual ratio of ions adjusted to neutralize the system. We follow the following protocol to equilibrate the structure: 1) Energy minimization (10,000 conjugate gradient steps); 2) Equilibration runs for 100 ns (NPT ensemble), where the temperature stays at 300 K. We record the coordinates for each 10 ps, compare with the PDB structure by RMSD to measure how far is the folded structure away from the PDB structure.

Peptide synthesis

The peptide used in this study was synthesized by GenScript (Piscataway, NJ), with free N- and C-termini. Peptides were synthetized using standard Fluorenylmethyloxycarbonyl (Fmoc)-based solid-phase peptide synthesis (SPPS) and purified by reverse-phase high-performance liquid chromatography (RP-HPLC) to a purity of 95% or higher.

Circular Dichroism (CD) Spectroscopy

Circular Dichroism (CD) spectra were recorded from 190 to 260 nm using a JASCO J-1500 spectrometer, with each spectrum averaged from three consecutive scans, the wavelength step being 0.5nm and the scan rate being 50 nm/min. Samples of 1mg/ml in deionized water were measured in a 0.1mm path length quartz cuvette (Starna Cells, Inc.). Secondary structure estimation was performed using the CONTINLL program with a reference set of 48 soluble proteins.

Fourier-transform Infrared Spectroscopy (FTIR)

ATR-FTIR measurements were performed on a Nicolet 6700 FT-IR spectrometer (Thermo Scientific) equipped with a liquid-nitrogen-cooled microscope. Spectra were collected in reflection mode with ATR correction using a germanium crystal. Each spectrum was collected from 4000 to 650 cm^{-1} with a resolution of 4 cm^{-1} and 64 scans. The relative fractions of different secondary structures were determined by Fourier self-deconvolution (FSD) of the Amide I band (1705-1595 cm^{-1}) and Gaussian curve-fitting of the deconvoluted spectra using Origin.

Raman Spectroscopy

Raman spectroscopy were performed using a Renishaw Invia Reflex Raman confocal microscope equipped with a 1” CCD array detector (1024 × 256 pixels), a 532 nm laser and a 100× objective. Each spectrum was collected in the 101 - 2736 cm^{-1} range and as an accumulation of 20 scans to increase the signal-to-noise ratio (3 seconds exposure per scan).

Transmission Electron Microscopy (TEM) imaging

Transmission electron micrographs were recorded using a Tecnai G^2 Spirit TWIN (LaB$_6$ filament, 120 kV) equipped with a Gatan CCD camera. Continuous-film carbon-coated copper grids (Ted Pella, CA) were glow discharged and used for negatively stained samples. Briefly, 5 µL peptide samples of 1 mg/ml in deionized water were pipetted onto the grid, wicked off after 2 minutes, washed with water and then stained with 5 μL 2% uranyl acetate for 1 minute before being wicked off. The grid was then left to dry before imaging. Dimensional measurements of the peptide assemblies were performed on ten different micrographs using DigitalMicrograph (Gatan Inc.).

Author contributions: The research was designed by MJB, BM, LW and PYC with input from all co-authors. Molecular modeling and analysis was conducted by ZQ and MJB. The AI work was led by LW, SH, TM and PYC. The paper was written with input from all co-authors. All co-authors approved this submission.

Acknowledgements: This research was supported by the IBM-MIT AI lab. The authors acknowledge the Center for Materials Science and Engineering (CMSE) and Biophysical Instrumentation Facility (BIF) at MIT for access to the structural characterization and imaging instruments.

References

1. Ebrahimi, D. *et al.* Silk-Its Mysteries, How It Is Made, and How It Is Used. *ACS Biomater. Sci. Eng.* **1**, (2015).

2. Gronau, G. *et al.* A review of combined experimental and computational procedures for assessing biopolymer structure-process-property relationships. *Biomaterials* **33**, (2012).

3. Herrmann, H. & Aebi, U. Intermediate Filaments: Molecular Structure, Assembly Mechanism, and Integration Into Functionally Distinct Intracellular Scaffolds. *Annu. Rev. Biochem.* (2004). doi:10.1146/annurev.biochem.73.011303.073823

4. Rowat, A. C., Lammerding, J., Herrmann, H. & Aebi, U. Towards an integrated understanding of the structure and mechanics of the cell nucleus. *BioEssays* (2008). doi:10.1002/bies.20720

5. Windoffer, R., Beil, M., Magin, T. M. & Leube, R. E. Cytoskeleton in motion: The dynamics of keratin intermediate filaments in epithelia. *J. Cell Biol.* (2011). doi:10.1083/jcb.201008095

6. Weisman, S. *et al.* Honeybee silk: Recombinant protein production, assembly and fiber spinning. *Biomaterials* (2010). doi:10.1016/j.biomaterials.2009.12.021

7. Sutherland, T. D. *et al.* Single honeybee silk protein mimics properties of multi-protein silk. *PLoS One* (2011). doi:10.1371/journal.pone.0016489

8. Sutherland, T. D., Sriskantha, A., Rapson, T. D., Kaehler, B. D. & Huttley, G. A. Did aculeate silk evolve as an antifouling material? *PLoS One* (2018). doi:10.1371/journal.pone.0203948

9. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **41**, 1–7 (2000).

10. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1049

11. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* (2010). doi:10.1038/nature09304

12. Conchúir, S. *et al.* A Web resource for standardized benchmark datasets, metrics, and rosetta protocols for macromolecular modeling and design. *PLoS One* (2015). doi:10.1371/journal.pone.0130433

13. Naganathan, A. N. & Muñoz, V. Scaling of folding times with protein size. *J. Am. Chem. Soc.* (2005). doi:10.1021/ja044449u

14. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding.

Chem. Phys. Lett. (1999). doi:10.1016/S0009-2614(99)01123-9

15. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* (2016). doi:10.1038/nature16961

16. Gu, G. X., Chen, C. T., Richmond, D. J. & Buehler, M. J. Bioinspired hierarchical composite design using machine learning: simulation, additive manufacturing, and experiment. *Mater. Horizons* **5**, 939–945 (2018).

17. Hanakata, P. Z., Cubuk, E. D., Campbell, D. K. & Park, H. S. Accelerated Search and Design of Stretchable Graphene Kirigami Using Machine Learning. *Phys. Rev. Lett.* (2018). doi:10.1103/PhysRevLett.121.255304

18. Gu, G. X., Chen, C. T. & Buehler, M. J. De novo composite design based on machine learning algorithm. *Extrem. Mech. Lett.* **18**, 19–28 (2018).

19. Yu, C. H., Qin, Z., Martin-Martinez, F. & Buehler, M. J. A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using AI. 7–9

20. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* (2018). doi:10.1126/sciadv.aap7885

21. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* (1983). doi:10.1002/bip.360221211

22. Burgess, N. C. *et al.* Modular Design of Self-Assembling Peptide-Based Nanotubes. *J. Am. Chem. Soc.* (2015). doi:10.1021/jacs.5b03973

23. Fletcher, J. M. *et al.* A basis set of de novo coiled-Coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* (2012). doi:10.1021/sb300028q

24. Thomson, A. R. *et al.* Computational design of water-soluble α-helical barrels. *Science (80-.).* (2014). doi:10.1126/science.1257452

25. Zaccai, N. R. *et al.* A de novo peptide hexamer with a mutable channel. *Nat. Chem. Biol.* (2011). doi:10.1038/nchembio.692

26. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* (2007). doi:10.1038/nprot.2006.202

27. Kelly, S. M., Jess, T. J. & Price, N. C. How to study proteins by circular dichroism. *Biochimica et Biophysica Acta - Proteins and Proteomics* (2005). doi:10.1016/j.bbapap.2005.06.005

28. Sreerama, N. & Woody, R. W. Estimation of protein secondary structure from circular dichroism spectra: Comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* (2000). doi:10.1006/abio.2000.4880

29. Hu, X., Kaplan, D. & Cebe, P. Determining beta-sheet crystallinity in fibrous proteins by thermal analysis and infrared spectroscopy. *Macromolecules* (2006). doi:10.1021/ma0610109

30. Hu, X., Kaplan, D. & Cebe, P. Dynamic protein-water relationships during β-sheet formation. *Macromolecules* (2008). doi:10.1021/ma071551d

31. Woodhead, A. L., Sutherland, T. D. & Church, J. S. Structural analysis of hand drawn bumblebee bombus terrestris silk. *Int. J. Mol. Sci.* (2016). doi:10.3390/ijms17071170

32. Lazaridis, T. & Karplus, M. '"New view"' of protein folding reconciled with the old through multiple unfolding simulations. *Science (80-.).* **278**, 1928–1931 (1997).

33. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* (1998). doi:10.1021/jp973084f

34. Nelson, M. T. *et al.* NAMD: A parallel, object-oriented molecular dynamics program. *Int. J. High Perform. Comput. Appl.* (1996). doi:10.1177/109434209601000401

35. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 27-28,33-38 (1996).

Tables

Table 1: This table summarizes the computational efficiency for different simulation and prediction methods and their error (RMSD as given by Eq. (1)), by comparing to the 3D structure as well as experimental results.

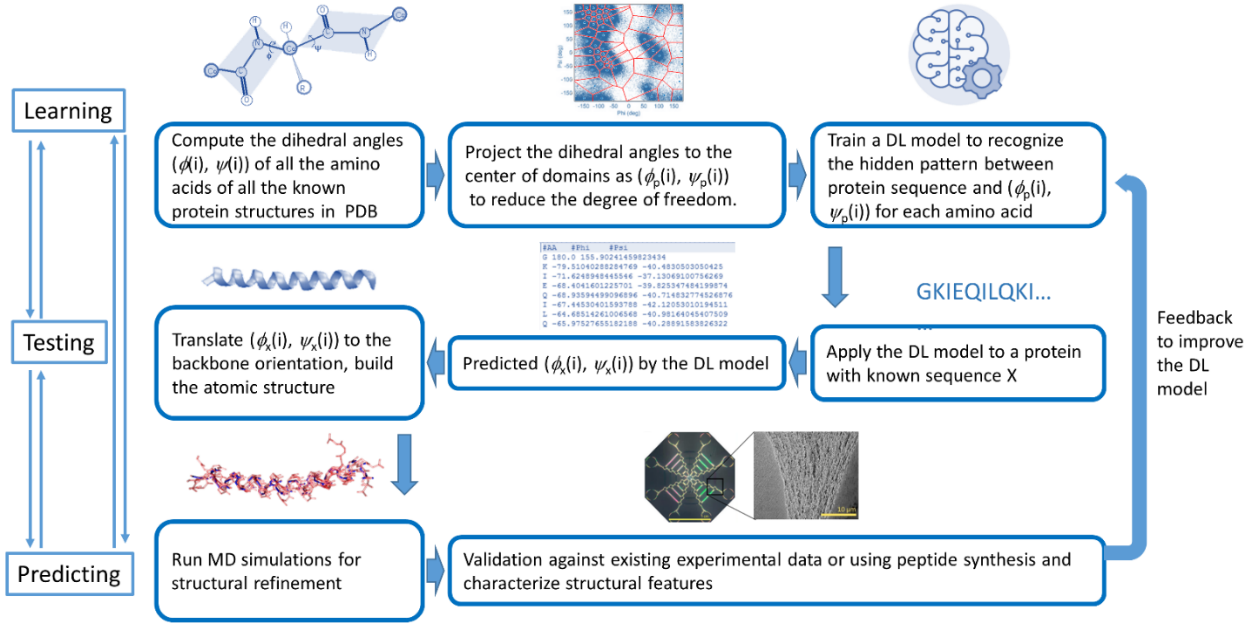
Sequence #	Straight+EEF (100 ns)*		MNNN +EEF (100 ns)*		Straight+Tip3 (100 ns)**		MNNN +Tip3 (100 ns)**		MNNN prediction ***	
	Time (hour)	Err$_{end}$ (Å)	Time (hour)	Err$_{end}$ (Å)	Time (hour)	Err$_{end}$ (Å)	Time (hour)	Err$_{end}$ (Å)	Time for prediction	Err (Å)
seq 1	38.3	14.7	38.4	11.5	46.9	8.6	115.5	1.8	0.65 sec	2.11
seq 2	41.0	12.6	43.0	11.9	121.6	10.8	113.1	1.8	0.65 sec	0.71
seq 3	41.0	11.1	41.4	8.7	121.3	9.8	113.9	1.4	0.65 sec	1.01
seq 4	39.9	11.0	39.4	10.5	121.4	7.0	32.3	1.5	0.65 sec	1.10
seq 5	46.2	11.2	43.2	10.6	121.5	10.5	37.3	0.9	0.65 sec	0.68
seq 6	36.3	13.8	38.5	10.4	114.8	10.6	32.7	1.1	0.65 sec	0.89

* computed by 1 Xeon CPU core

** computed by 4 Xeon CPU cores + GPU (one of Nvidia 1070, 1080 or 1080 Ti)

*** computed by 1 i7 CPU

Figures



Learning

Testing

Predicting

Compute the dihedral angles (ϕ(i), ψ(i)) of all the amino acids of all the known protein structures in PDB

Project the dihedral angles to the center of domains as (ϕ_p(i), ψ_p(i)) to reduce the degree of freedom.

Train a DL model to recognize the hidden pattern between protein sequence and (ϕ_p(i), ψ_p(i)) for each amino acid

Apply the DL model to a protein with known sequence X

GKIEQILQKI...

Predicted (ϕ_x(i), ψ_x(i)) by the DL model

```
#AA   #Phi    #Psi
G 180.0 155.90241459823434
K -79.51040288284769 -40.4830503050425
I -71.6248948445546 -37.13069100756269
E -68.4041601225701 -39.825347484199874
Q -68.93594499096896 -40.714832774526876
I -67.44530401593788 -42.1205301019451I
L -64.68514261006568 -40.98164045407509
Q -65.97527655182188 -40.28891583826322
```

Translate (ϕ_x(i), ψ_x(i)) to the backbone orientation, build the atomic structure

Run MD simulations for structural refinement

Validation against existing experimental data or using peptide synthesis and characterize structural features

Feedback to improve the DL model

Figure 1: Overall flowchart of the algorithm reported in this paper. Taking the entire Protein Data Bank (composed of ~120,000 protein structures) as the training set, we extract the sequence information, along with the phi-psi angle information from each Protein Data Bank file for the high-resolution protein structure. We further label the dihedral angles by considering the natural distribution of phi-psi angle to reduce the degree of freedom and use the structural labels, combined with the sequence information to train a MNNN model, which allows us to predict the phi-psi angle of any sequence and thus build the atomic structure together with the other intrinsic coordinate parameters. We run long-time MD simulations to quantify the acceleration of MNNN prediction and the stability of the structure given by MNNN result. The result is compared with experimental synthesis and characterization, which provide feedback to improve the quality of the MNNN model.

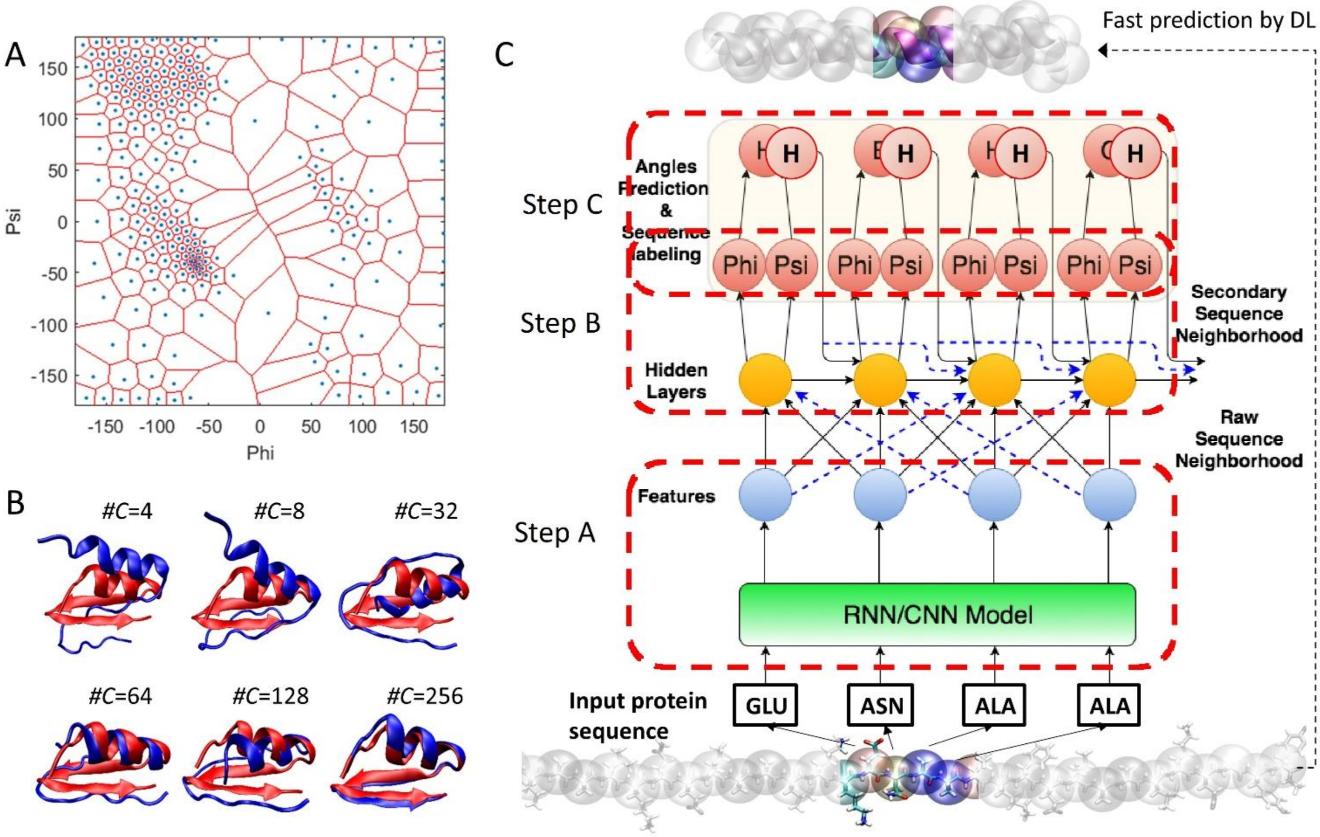
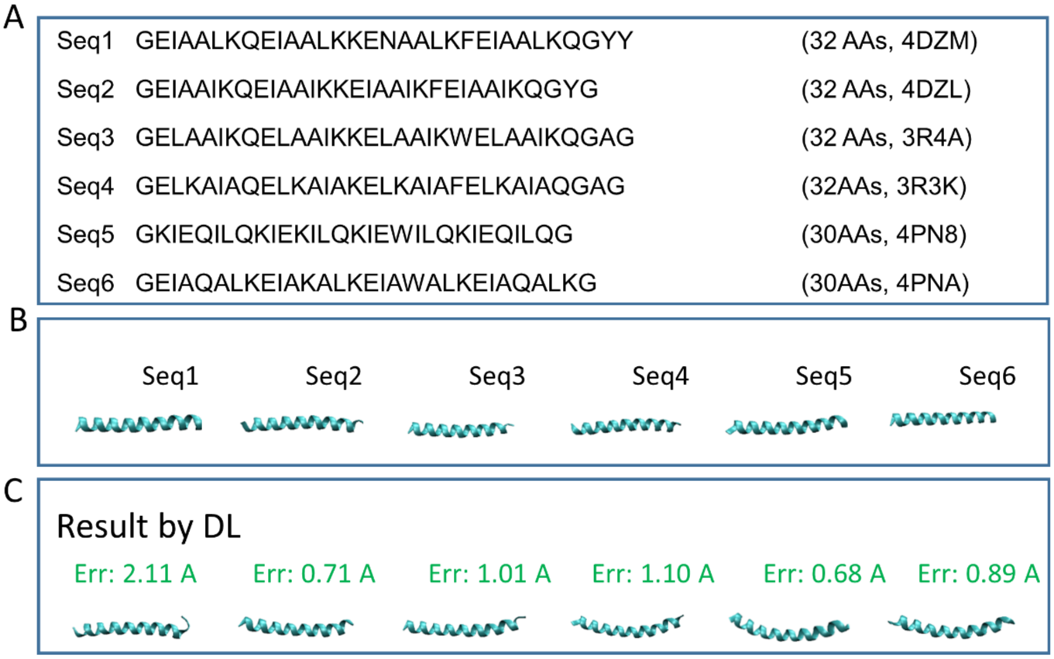


Figure 2: Strategy to reduce the design space of backbone conformations while achieving a high fidelity and overall architecture of our proposed MNNN model. A) We use a K means clustering algorithm to categorize all phi-psi angles in PDB into 256 clusters, which effectively reduces the infinite combination of phi-psi angles to the value of one of the cluster center (0..255). B) This panel shows the performance of using a different number of clusters in representing a protein structure (blue color, PDB ID: 1ACW), in comparison with the PDB structure (red). It is shown that for a 256 cluster choice, the error is reduced to a small level. C) The architecture of our MNNN model (step A, B, and C) takes into account both information of raw sequence neighborhood and of secondary sequence neighborhood. In step A, we compute character embedding for each amino acid using any popular techniques and then refine these embeddings with sequence neighborhood information. In step B, we apply another LSTM layer to perform dihedral angles prediction based on refined character embeddings of amino acid sequence. In Step C, we use these two embeddings to further predict secondary structure characters as additional constraints based on predicted dihedral angles.



Seq1	GEIAALKQEIAALKKENAALKFEIAALKQGYY	(32 AAs, 4DZM)
Seq2	GEIAAIKQEIAAIKKEIAAIKFEIAAIKQGYG	(32 AAs, 4DZL)
Seq3	GELAAIKQELAAIKKELAAIKWELAAIKQGAG	(32 AAs, 3R4A)
Seq4	GELKAIAQELKAIAKELKAIAFELKAIAQGAG	(32AAs, 3R3K)
Seq5	GKIEQILQKIEKILQKIEWILQKIEQILQG	(30AAs, 4PN8)
Seq6	GEIAQALKEIAKALKEIAWALKEIAQALKG	(30AAs, 4PNA)

Figure 3: Summary of all sequences investigated for benchmarking, and a brief summary of the prediction results. A) The six sequences and their corresponding structure id in PDB. B) The snapshots of the protein structures as obtained from the PDB. C) The result of MNNN and the RMSD value from PDB structures given in panel B.

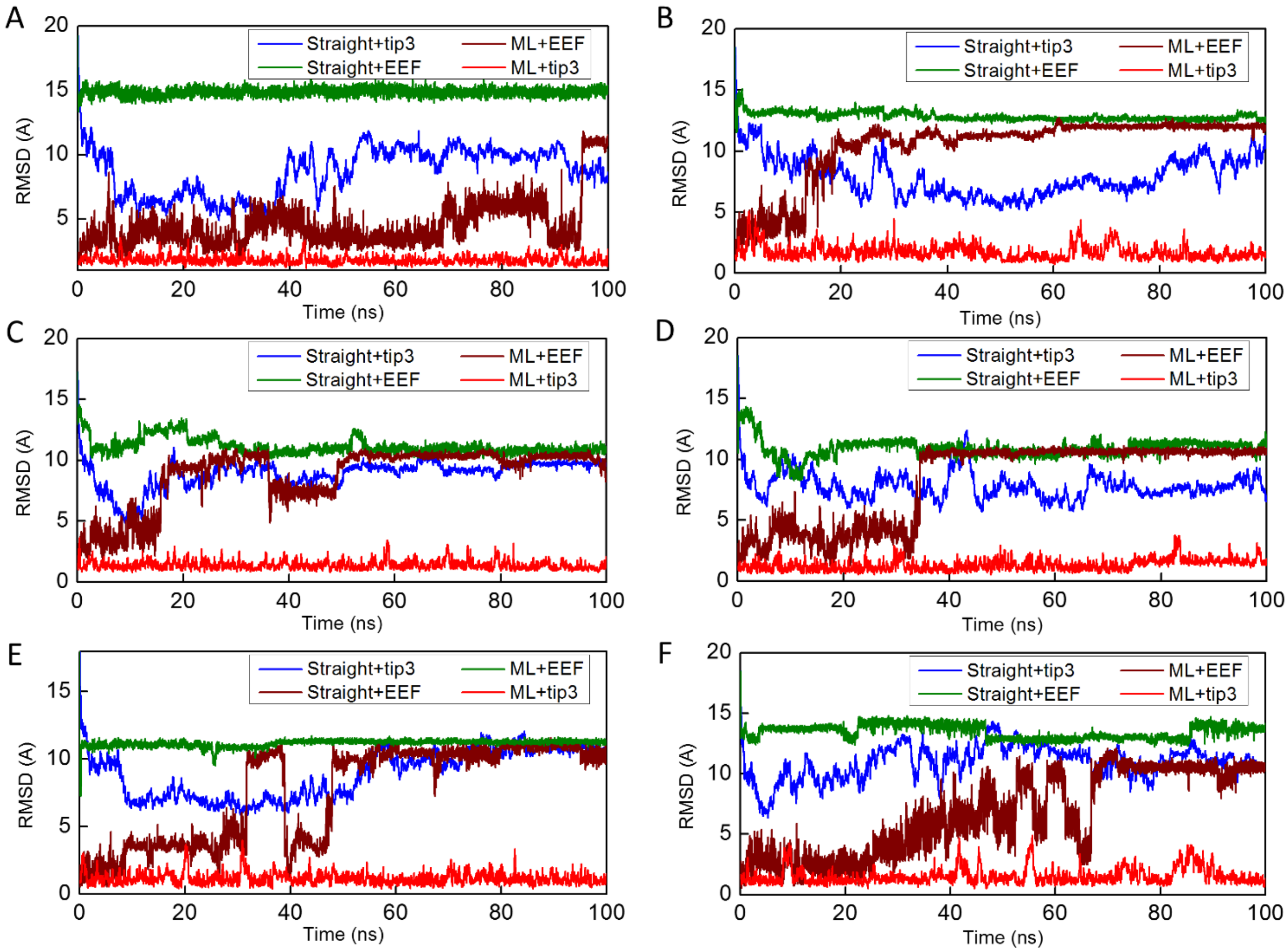


Figure 4: Benchmark results of all six sequences as summarized in Fig. 4. Each of them include different starting conformations (arbitrary straight chain and result obtained by MNNN) and different force fields and solvent models (Tip3P explicit solvent and EEF1 implicit solvent), with panels (A), (B), (C), (D), (E) and (F) for protein 4DZM, 4DZL, 3R4A, 3R3K, 4PN8 and 4PNA, respectively.

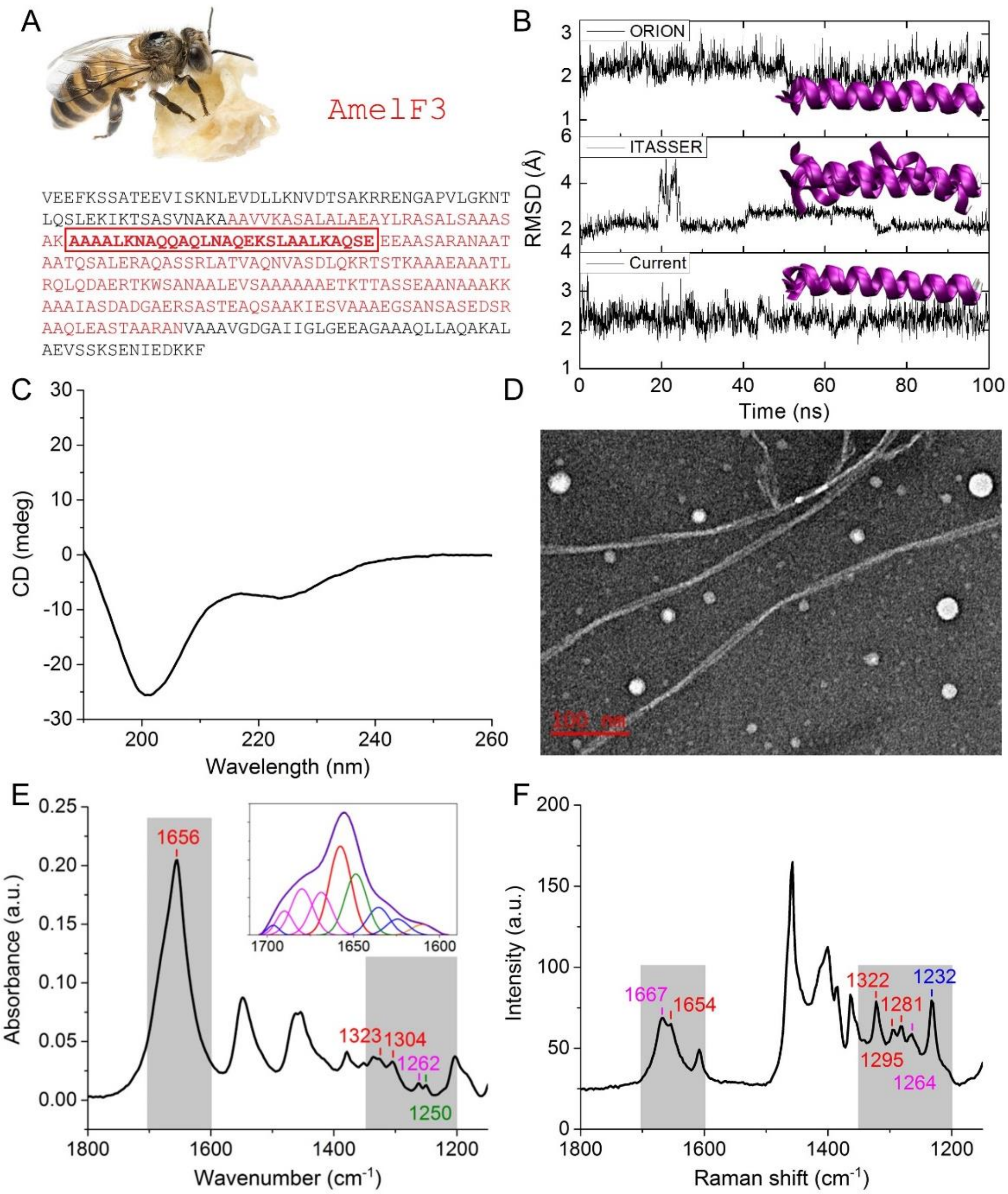


Figure 5: Synthesis and characterization of a *de novo* peptide sequence that is not part of the Protein Data Bank. A) Amino acid sequence of honeybee silk protein AmelF3, with the coiled-coil domain highlighted in red and the peptide (AmelF3_+1) sequence shown in **bold**. B) Results of 100 ns MD simulation starting from the predictions given by ORION, I-TASSER and the MNNN model. Snapshots taken every 20 ns of each simulation are overlaid for comparison. C) CD spectrum of AmelF3_+1 in deionized water. D) Representative transmission electron micrograph of the AmelF3_+1 peptide, depicting nanofiber formation. E) FTIR spectrum of AmelF3_+1 film, with alpha-helical peaks labeled in red, turns in magenta, and random coils in olive. The inset shows Fourier self-deconvolution and secondary structure peak fitting of the Amide I band. F) Raman spectrum of AmelF3_+1 film, with the peaks labeled by the same color notation for secondary structures as used in the FTIR spectrum, blue represents beta-sheet.